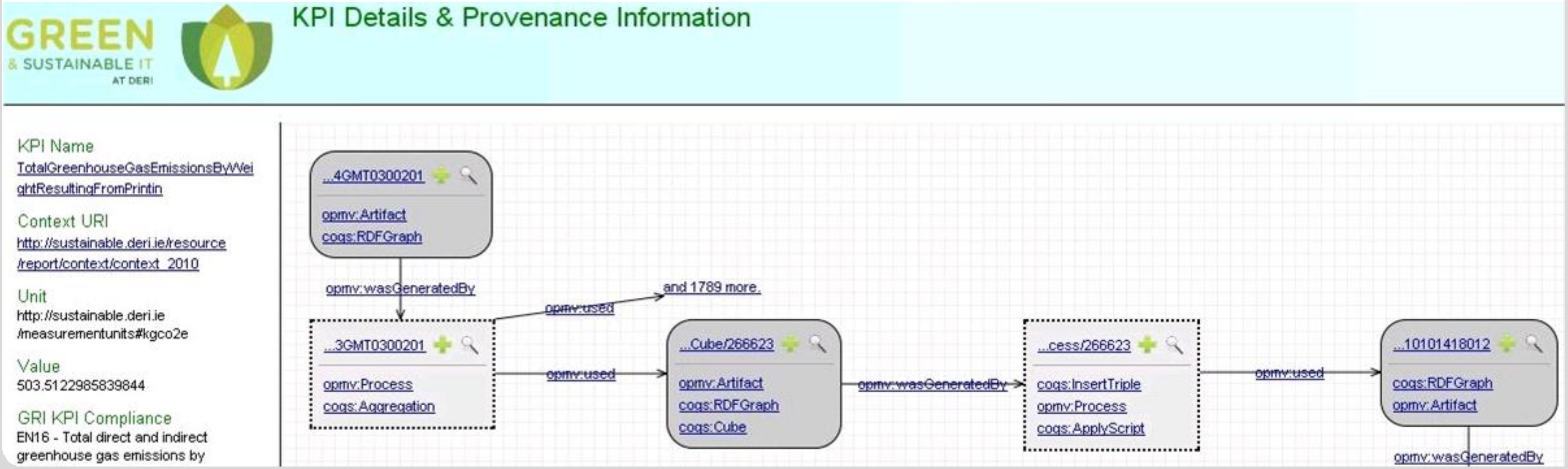
Representing Interoperable Provenance Descriptions for ETL Workflows

André Freitas, Benedikt Kämpgen, Joao Gabriel Oliveira, Seán O'Riain, Edward Curry

The role of Semantic Web in Provenance Management, Extended Semantic Web Conference 2012

28 May 2012

Institute of Applied Informatics and Formal Description Methods (AIFB)



GREEN & SUSTAINABLE IT AT DERI

KPI Details & Provenance Information

KPI Name
TotalGreenhouseGasEmissionsByWeightResultingFromPrintin

Context URI
http://sustainable.deri.ie/resource/report/context/context_2010

Unit
http://sustainable.deri.ie/measurementunits#kgco2e

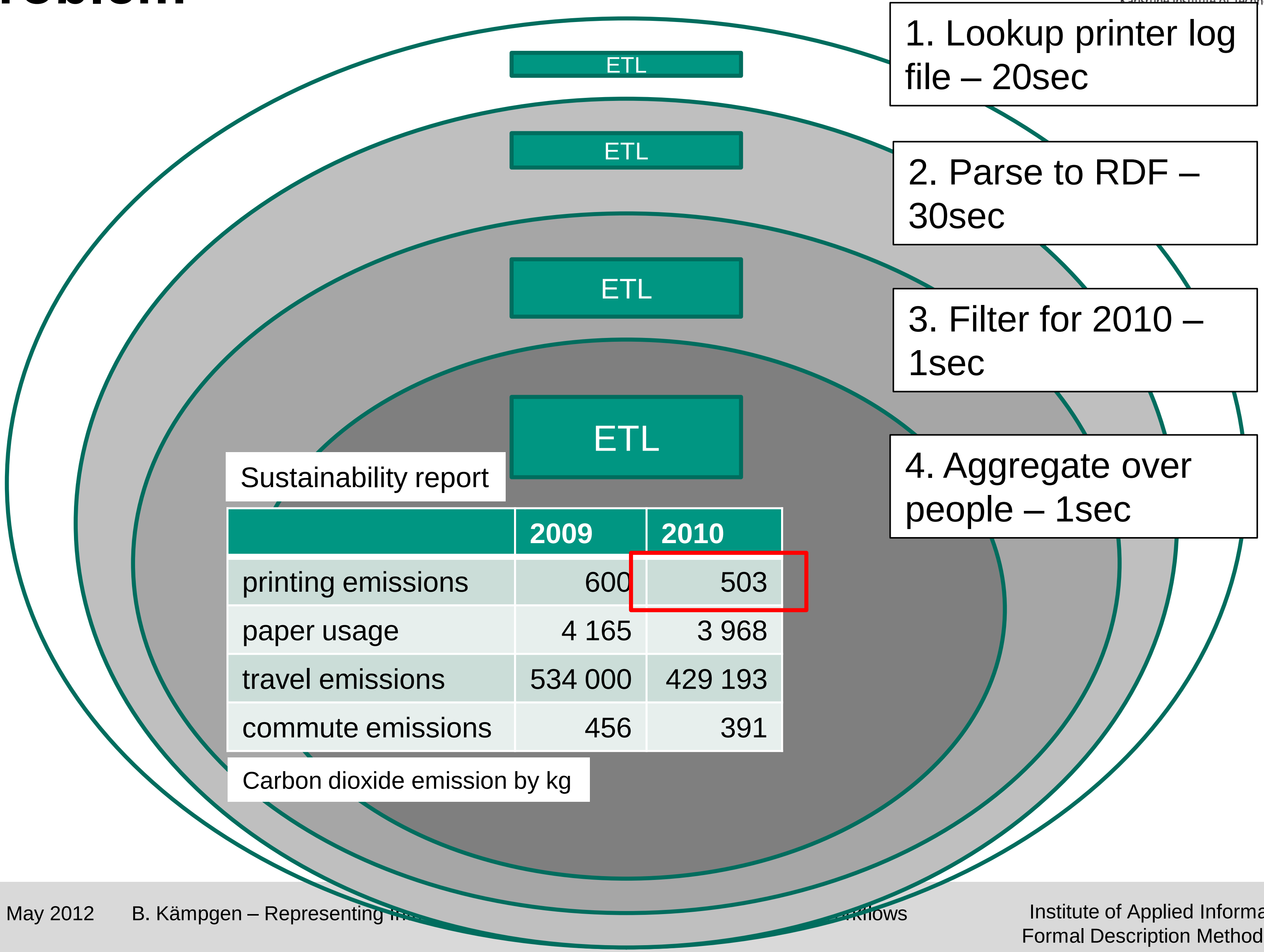
Value
503.5122985839844

GRI KPI Compliance
EN16 - Total direct and indirect greenhouse gas emissions by

Motivation

- **Decision-support** on more complex and heterogeneous data environments (dataspaces, Linked Open Data)
- **Extract-Transform-Load** (ETL) workflows inherent part of data analysis
- **Challenges**:
  - Management of complex ETL workflows
  - Information quality, trust

Problem



1. Lookup printer log file – 20sec

2. Parse to RDF – 30sec

3. Filter for 2010 – 1sec

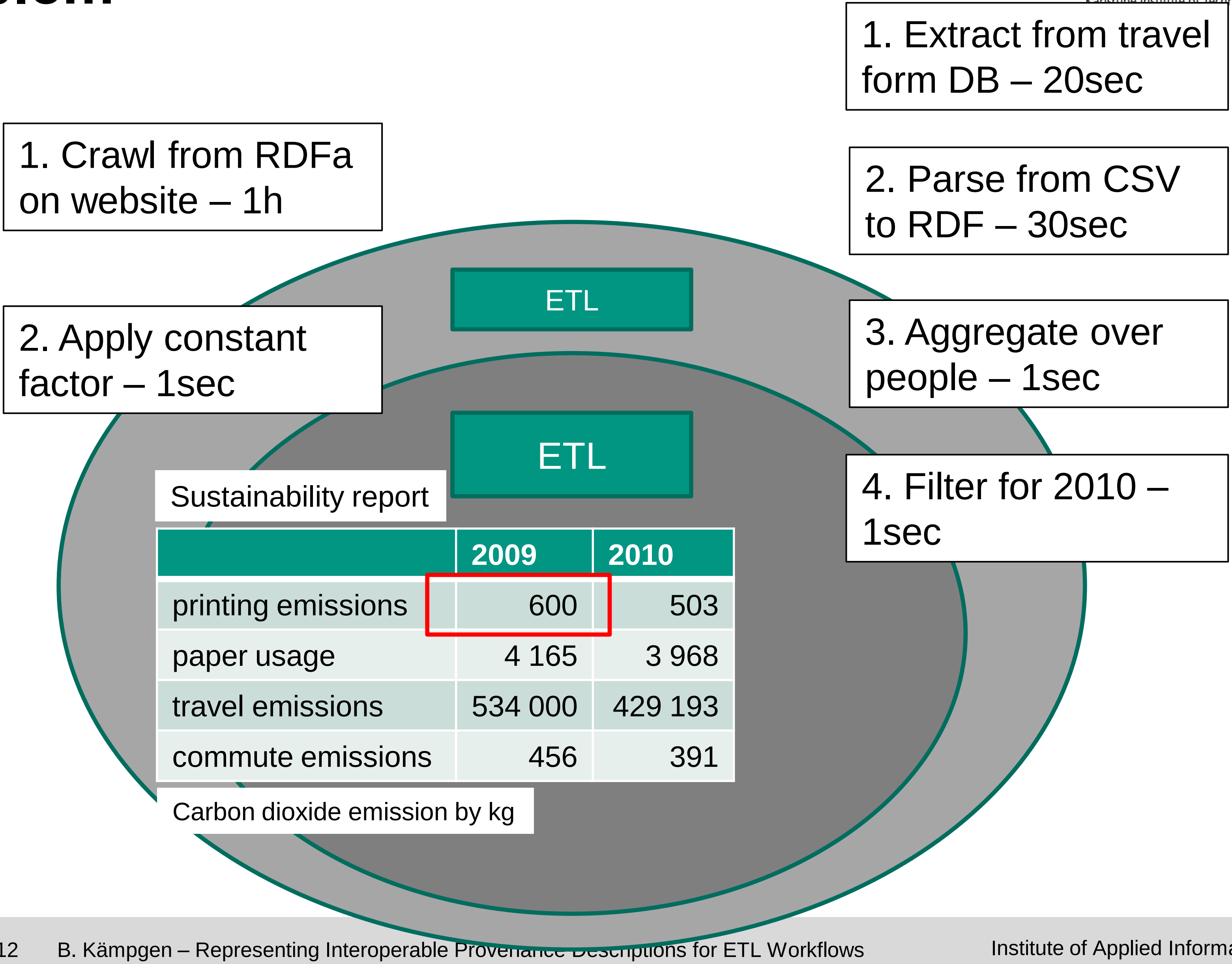
4. Aggregate over people – 1sec

Sustainability report

	2009	2010
printing emissions	600	503
paper usage	4 165	3 968
travel emissions	534 000	429 193
commute emissions	456	391

Carbon dioxide emission by kg

Problem



1. Crawl from RDFa on website – 1h

2. Apply constant factor – 1sec

1. Extract from travel form DB – 20sec

2. Parse from CSV to RDF – 30sec

3. Aggregate over people – 1sec

4. Filter for 2010 – 1sec

Sustainability report

	2009	2010
printing emissions	600	503
paper usage	4 165	3 968
travel emissions	534 000	429 193
commute emissions	456	391

Carbon dioxide emission by kg

Solution: Provenance information about ETL workflows

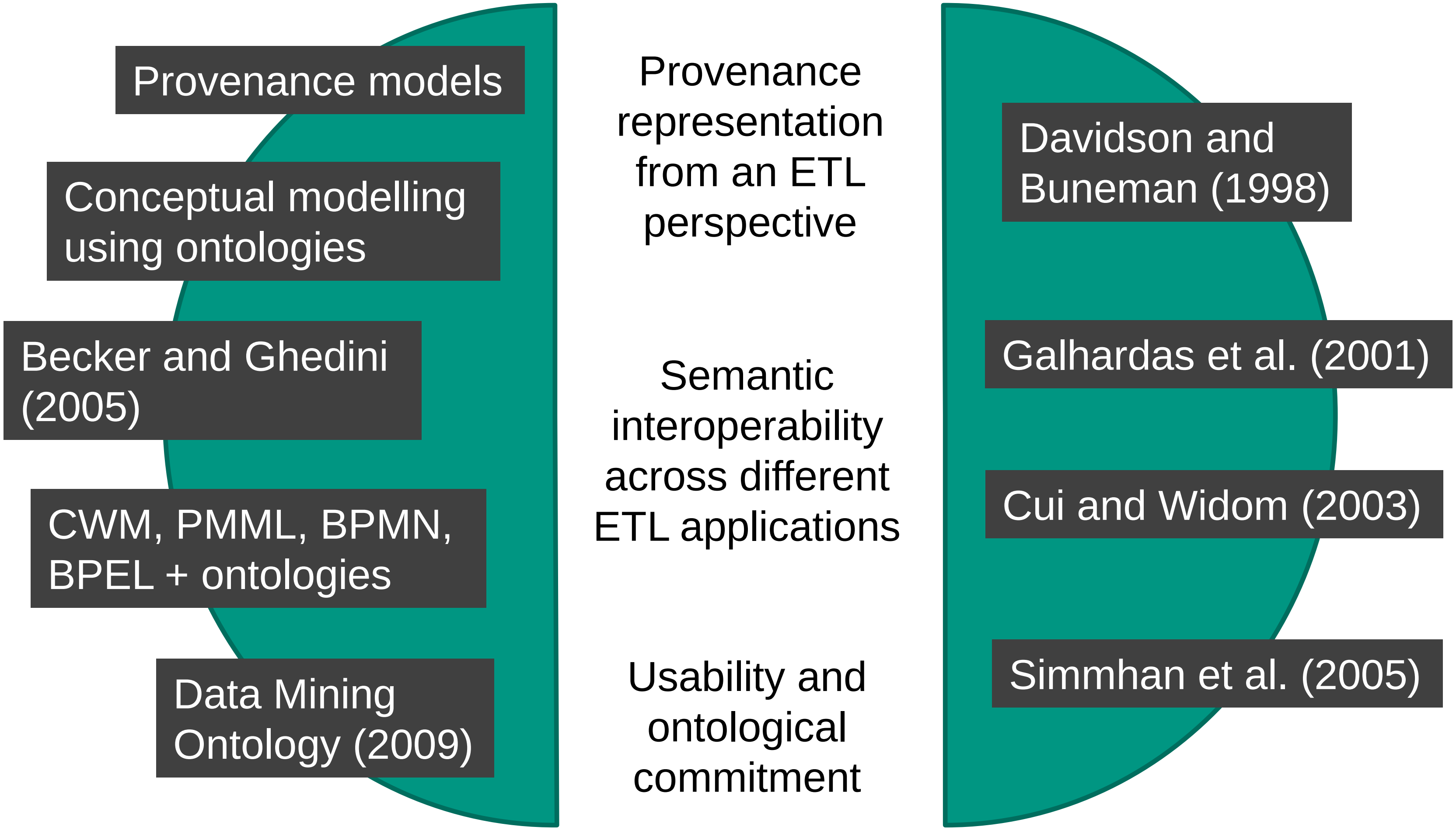
- Prospective provenance: representation of ETL workflow at design time
- Retrospective provenance: representation of ETL workflow after execution

- Applications of provenance information for ETL workflows
  - Documentation (reproducibility and reuse)
  - Data quality assessment (trustworthiness)
  - Management (consistency-checking, debugging and semantic reconciliation)

Outline

- Motivation & Problem
- **Gap of ETL Descriptions**
- Interoperable ETL Provenance Model
- Case Study
- Conclusions

Gap of ETL Descriptions (1)



Provenance models

Conceptual modelling using ontologies

Becker and Ghedini (2005)

CWM, PMML, BPMN, BPEL + ontologies

Data Mining Ontology (2009)

Provenance representation from an ETL perspective

Semantic interoperability across different ETL applications

Usability and ontological commitment

Davidson and Buneman (1998)

Galhardas et al. (2001)

Cui and Widom (2003)

Simmhan et al. (2005)

Conceptual modelling of ETL workflows

Interoperable ETL provenance model

Formal models of ETL workflows

Gap of ETL Descriptions (2)

- Common ETL applications
  - such as **Kapow Software**, **Pentaho Data Integration**, **Google Refine** and **Yahoo Pipes**
  - do not create and use provenance information or
  - do not support sharing and integrating such provenance information

Outline

- Motivation & Problem
- Gap of ETL Descriptions
- **Interoperable ETL Provenance Model**
- Case Study
- Conclusions

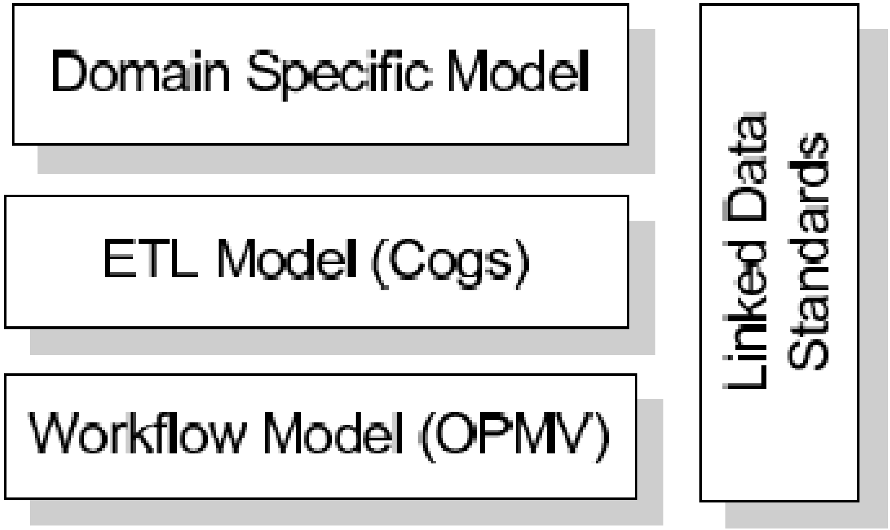
Outline

- Motivation & Problem
- Gap of ETL Descriptions
- **Interoperable ETL Provenance Model**
    - Requirements Analysis
    - High-level approach
    - Cogs: Linked Data vocabulary
    - Requirements Coverage Analysis
- Case Study
- Conclusions

Requirements Analysis

	Provenance representation from an ETL perspective	Semantic interoperability across different ETL platforms	Usability and ontological commitment
Prospective and retrospective descriptions	+	+	
Separation of concerns		+	
Common terminology	+	+	
Terminological completeness	+	+	
Lightweight ontology structure			+
Availability of different abstraction levels		+	+
Data representation independency			+
Accessibility		+	+
Decentralization		+	+

Interoperable Provenance Model for ETL Workflows

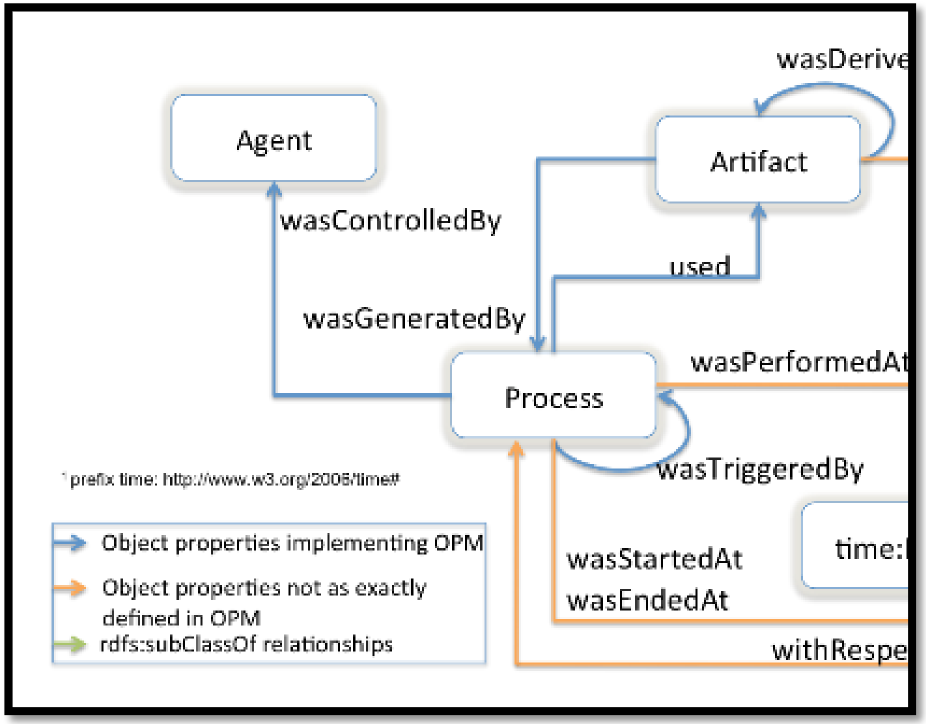


Three-layered Provenance Model

- High-level approach
  - reuse of the OPM Vocabulary (OPMV) workflow structure as abstract provenance model
  - creation of Cogs, an RDF vocabulary for representing ETL Provenance
  - can be extended by domain specific models
  - use of the Linked Data principles for representing provenance descriptors

Open Provenance Model Vocabulary (OPMV)

- Community-built provenance model
- Simple workflow structure (processes, artifacts, agents)
- Designed to be a minimal level of provenance interoperability
- Designed to be extensible
- ETL and provenance share workflow-level semantics


wasDerive
Agent
Artifact
wasControlledBy
used
wasGeneratedBy
wasPerformedAt
Process
wasTriggeredBy
¹ prefix time: http://www.w3.org/2006/time#
Object properties implementing OPM
Object properties not as exactly defined in OPM
rdfs:subClassOf relationships
wasStartedAt
wasEndedAt
time:I
withRespe

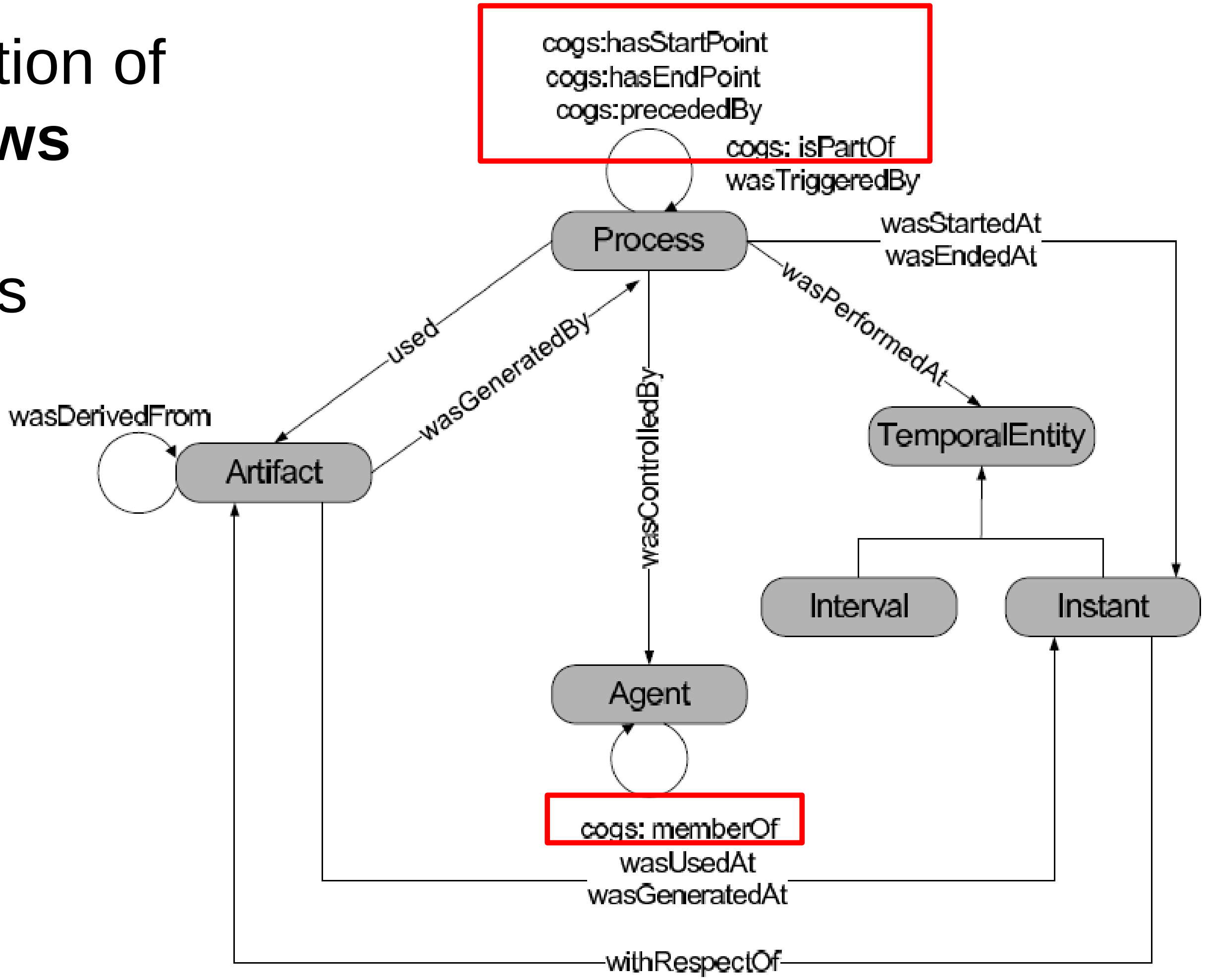
http://open-biomed.sourceforge.net/opmv/ns.html

Cogs

- RDF vocabulary for representing ETL elements
- Complementary vocabulary for expressing the elements present in an ETL workflow based on
  - ETL/data transformation tools (Pentaho Data Integration, Google Refine)
  - Concepts and structures from the ETL literature.
- https://sites.google.com/site/cogsvocab/

Cogs – OPMV workflow extension

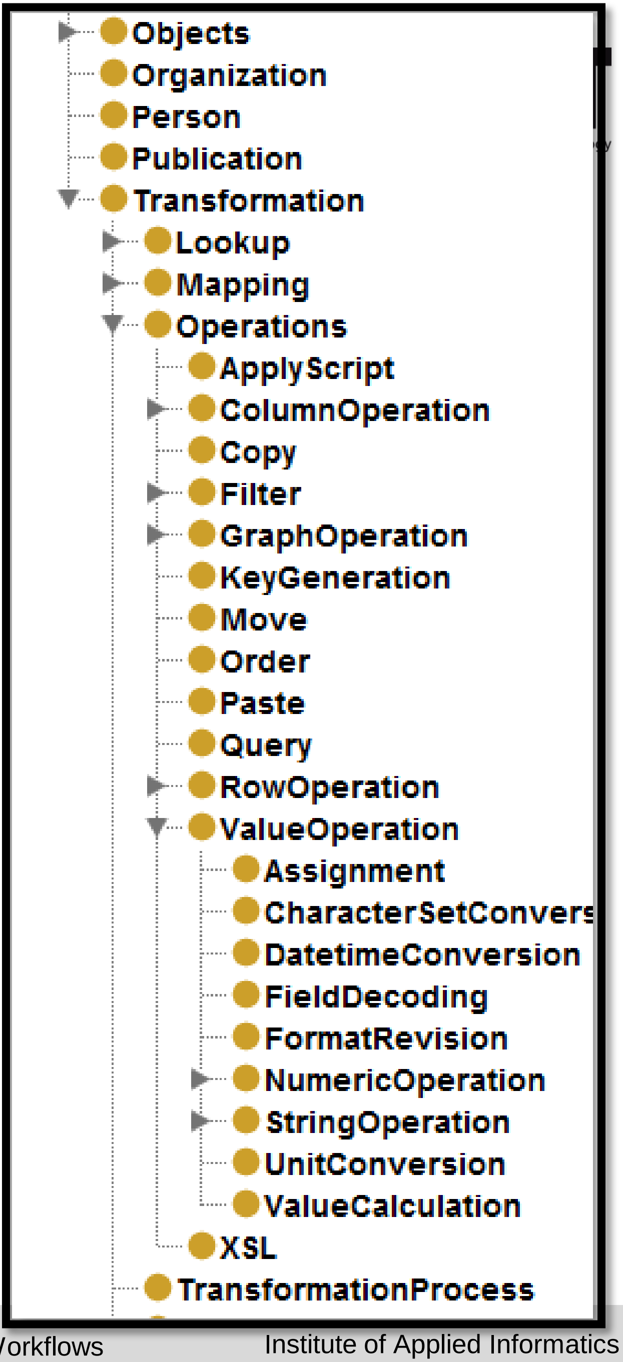
- The representation of **nested workflows** allows different abstraction levels



Cogs – Structure

- Taxonomy of ETL elements mapping to provenance processes and artifacts
- High-level classes:
  - cogs:Execution, e.g., ScheduledJob
  - cogs:State, e.g., Running
  - opmv:Process
    - cogs:Extraction, e.g., Parsing
    - cogs:Transformation, e.g., RegexFilter
    - cogs:Loading, e.g., IncrementalLoad
  - opmv:Artifact
    - cogs:Object, e.g., CSV File
  - cogs:Layer, e.g., StagingArea

Cogs:
151 classes
17 properties



- Objects
- Organization
- Person
- Publication
- Transformation
  - Lookup
  - Mapping
  - Operations
    - ApplyScript
    - ColumnOperation
    - Copy
    - Filter
    - GraphOperation
    - KeyGeneration
    - Move
    - Order
    - Paste
    - Query
    - RowOperation
    - ValueOperation
      - Assignment
      - CharacterSetConvers
      - DatetimeConversion
      - FieldDecoding
      - FormatRevision
      - NumericOperation
      - StringOperation
      - UnitConversion
      - ValueCalculation
    - XSL
  - TransformationProcess

Requirements Coverage Analysis

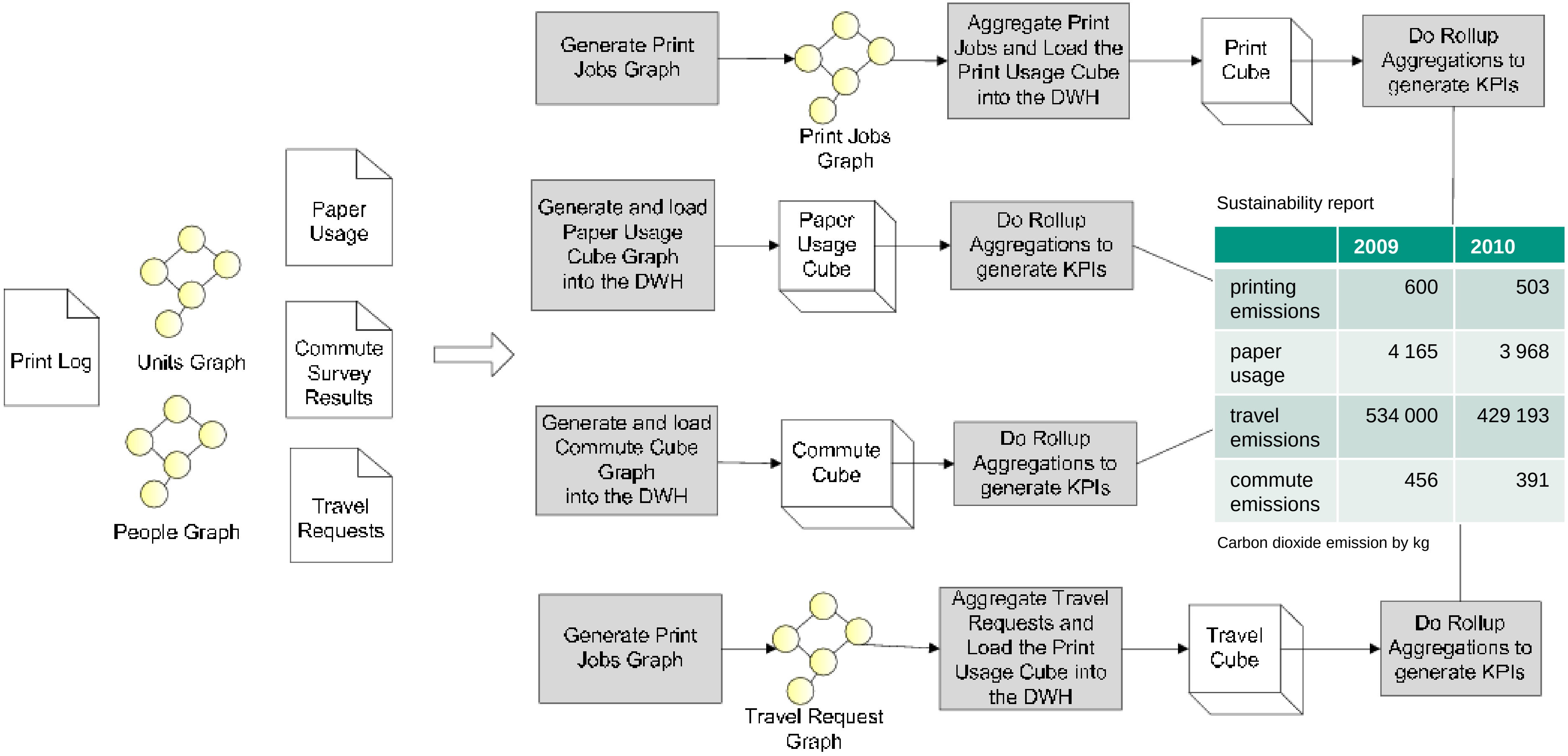
	OPMV	Cogs	LD principles
Prospective and retrospective descriptions	+	+	
Separation of concerns	+	+	
Common terminology	+	+	
Terminological completeness	+	+	+
Lightweight ontology structure	+	+	
Availability of different abstraction levels		+	
Data representation independency	+	+	+
Accessibility	+		+
Decentralization			+

Outline

- Motivation & Problem
- Gap of ETL Descriptions
- Interoperable ETL Provenance Model
- **Case Study**
- Conclusions

Case Study – Sustainability Reporting

- ETL over heterogeneous data sources (e.g., log files, survey results, travel request DB, RDF)


Print Log
Units Graph
People Graph
Paper Usage
Commute Survey Results
Travel Requests
Generate Print Jobs Graph
Print Jobs Graph
Aggregate Print Jobs and Load the Print Usage Cube into the DWH
Print Cube
Do Rollup Aggregations to generate KPIs
Generate and load Paper Usage Cube Graph into the DWH
Paper Usage Cube
Do Rollup Aggregations to generate KPIs
Generate and load Commute Cube Graph into the DWH
Commute Cube
Do Rollup Aggregations to generate KPIs
Generate Print Jobs Graph
Travel Request Graph
Aggregate Travel Requests and Load the Print Usage Cube into the DWH
Travel Cube
Do Rollup Aggregations to generate KPIs

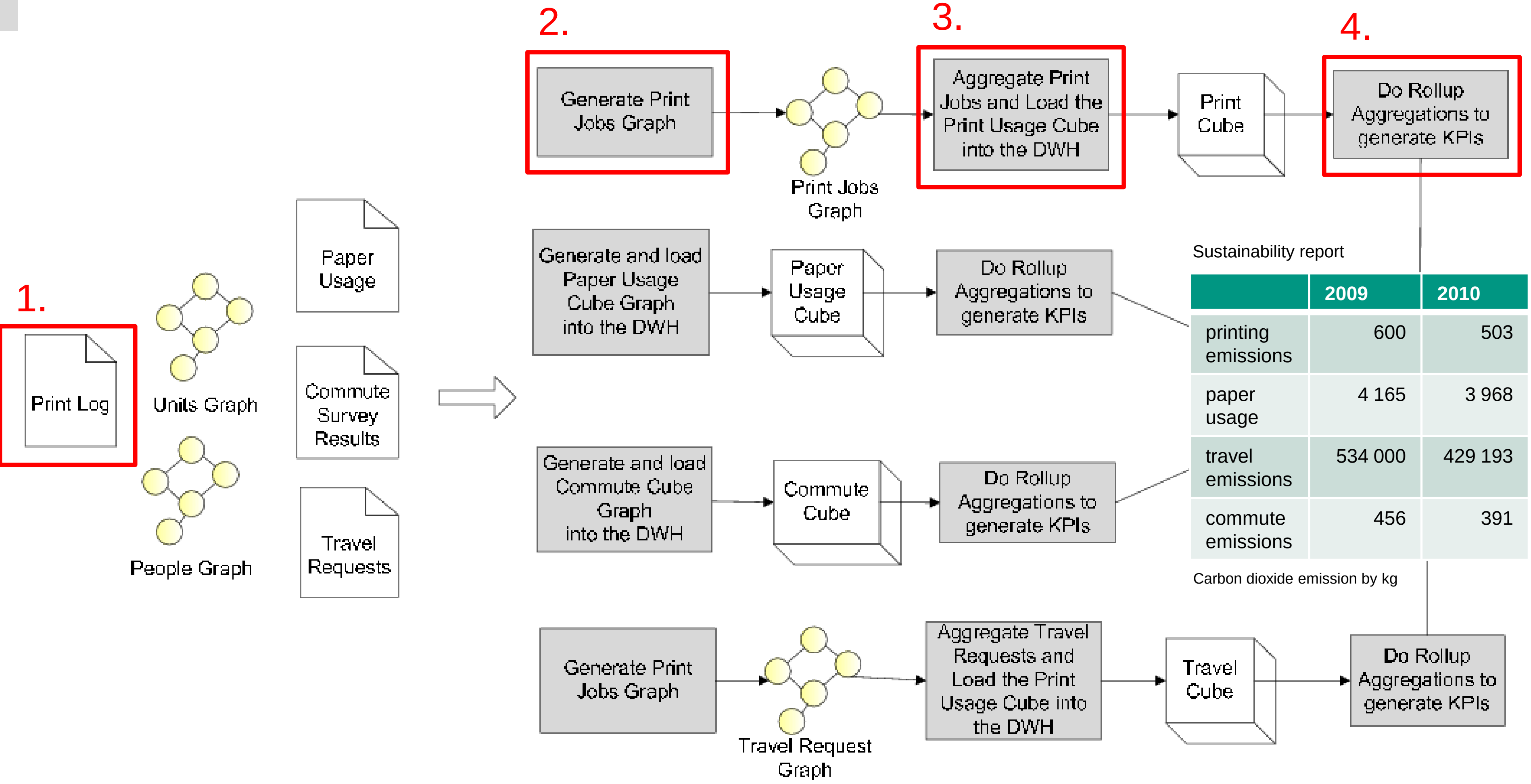
Sustainability report

	2009	2010
printing emissions	600	503
paper usage	4 165	3 968
travel emissions	534 000	429 193
commute emissions	456	391

Carbon dioxide emission by kg

Case Study – Sustainability Reporting

- ETL over heterogeneous data sources (e.g., log files, survey results, travel request DB, RDF)

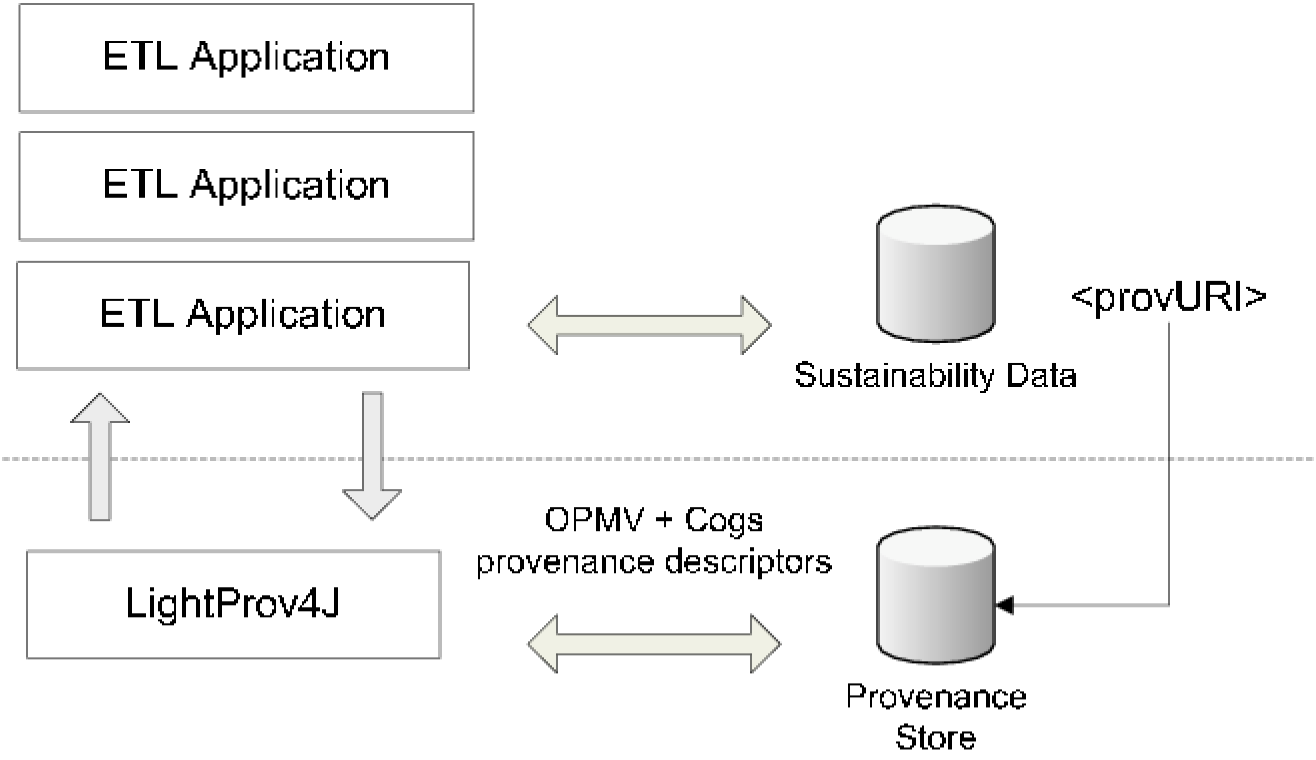

1.
Print Log
Units Graph
People Graph
Paper Usage
Commute Survey Results
Travel Requests
2.
Generate Print Jobs Graph
Print Jobs Graph
3.
Aggregate Print Jobs and Load the Print Usage Cube into the DWH
Print Cube
4.
Do Rollup Aggregations to generate KPIs
Generate and load Paper Usage Cube Graph into the DWH
Paper Usage Cube
Do Rollup Aggregations to generate KPIs
Generate and load Commute Cube Graph into the DWH
Commute Cube
Do Rollup Aggregations to generate KPIs
Generate Print Jobs Graph
Travel Request Graph
Aggregate Travel Requests and Load the Print Usage Cube into the DWH
Travel Cube
Do Rollup Aggregations to generate KPIs

Sustainability report

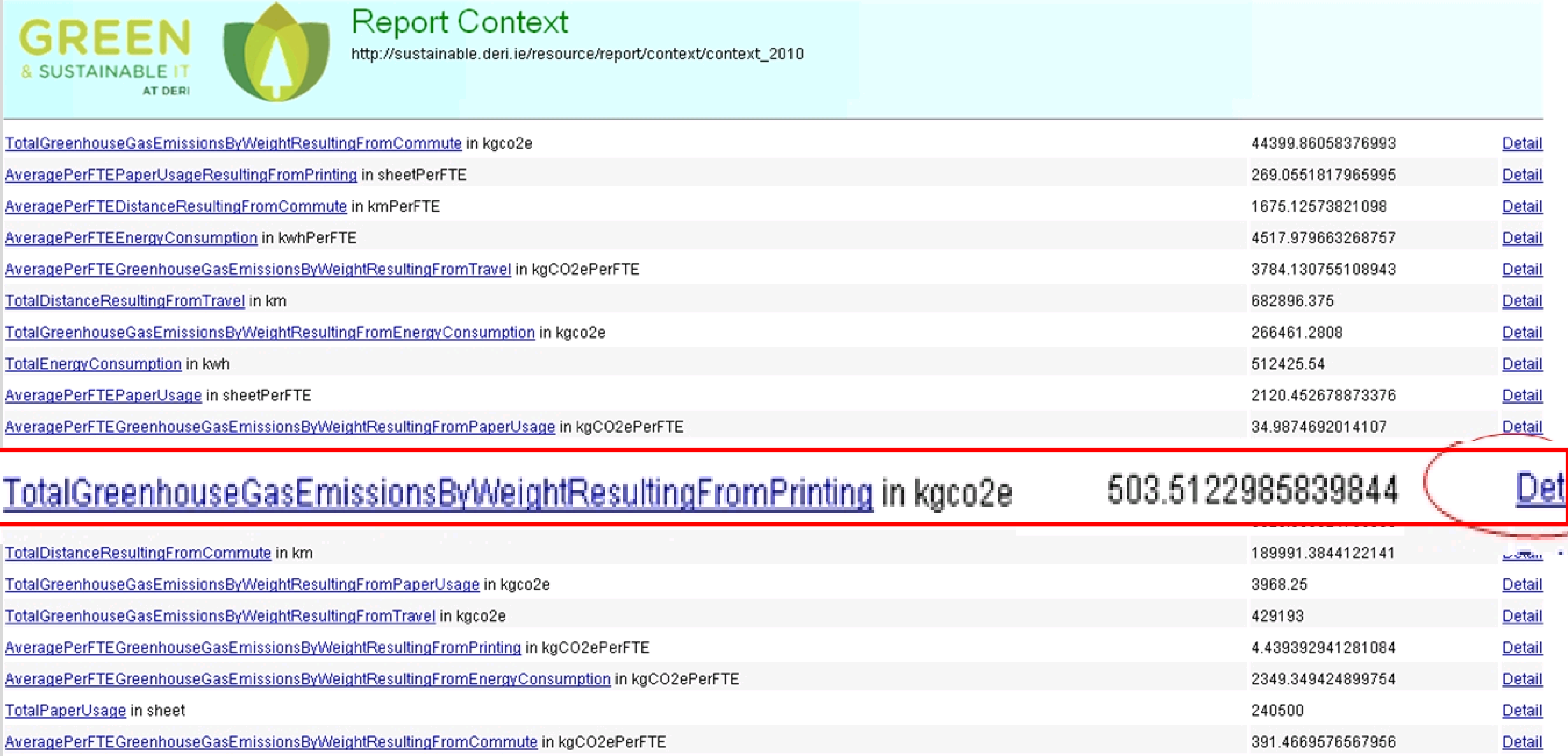
	2009	2010
printing emissions	600	503
paper usage	4 165	3 968
travel emissions	534 000	429 193
commute emissions	456	391

Carbon dioxide emission by kg

Case Study – Architecture with Provenance-aware ETL Applications



Case Study – Sustainability Report Values



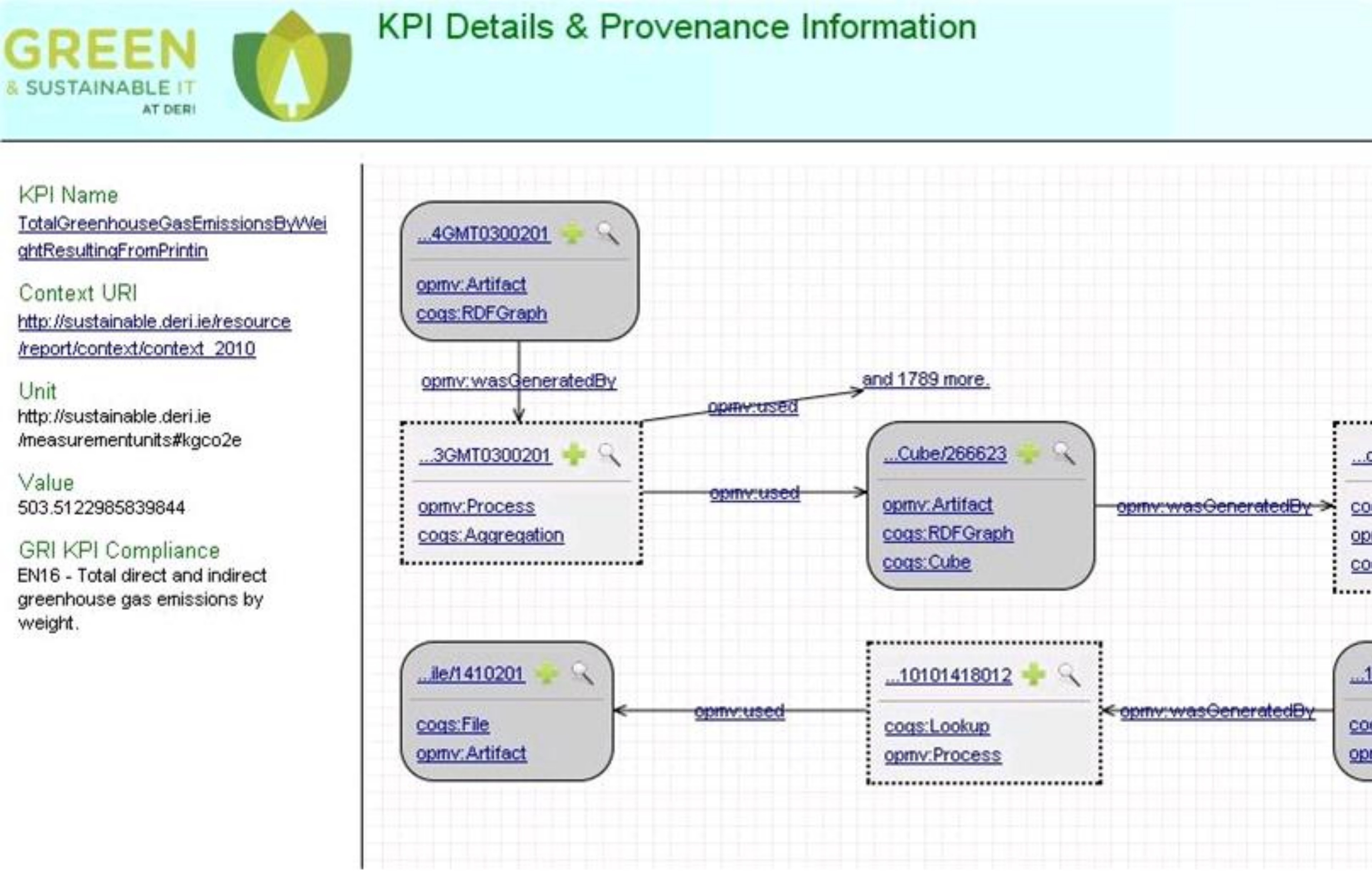
GREEN & SUSTAINABLE IT AT DERI

Report Context

http://sustainable.deri.ie/resource/report/context/context_2010

TotalGreenhouseGasEmissionsByWeightResultingFromCommute in kgco2e	44399.86058376993	Detail
AveragePerFTEPaperUsageResultingFromPrinting in sheetPerFTE	269.0551817965995	Detail
AveragePerFTEDistanceResultingFromCommute in kmPerFTE	1675.12573821098	Detail
AveragePerFTEEnergyConsumption in kwhPerFTE	4517.979663268757	Detail
AveragePerFTEGreenhouseGasEmissionsByWeightResultingFromTravel in kgCO2ePerFTE	3784.130755108943	Detail
TotalDistanceResultingFromTravel in km	682896.375	Detail
TotalGreenhouseGasEmissionsByWeightResultingFromEnergyConsumption in kgco2e	266461.2808	Detail
TotalEnergyConsumption in kwh	512425.54	Detail
AveragePerFTEPaperUsage in sheetPerFTE	2120.452678873376	Detail
AveragePerFTEGreenhouseGasEmissionsByWeightResultingFromPaperUsage in kgCO2ePerFTE	34.9874692014107	Detail
TotalGreenhouseGasEmissionsByWeightResultingFromPrinting in kgco2e	503.5122985839844	Det
TotalDistanceResultingFromCommute in km	189991.3844122141	Detail
TotalGreenhouseGasEmissionsByWeightResultingFromPaperUsage in kgco2e	3968.25	Detail
TotalGreenhouseGasEmissionsByWeightResultingFromTravel in kgco2e	429193	Detail
AveragePerFTEGreenhouseGasEmissionsByWeightResultingFromPrinting in kgCO2ePerFTE	4.439392941281084	Detail
AveragePerFTEGreenhouseGasEmissionsByWeightResultingFromEnergyConsumption in kgCO2ePerFTE	2349.349424899754	Detail
TotalPaperUsage in sheet	240500	Detail
AveragePerFTEGreenhouseGasEmissionsByWeightResultingFromCommute in kgCO2ePerFTE	391.4669576567956	Detail

Case Study – Provenance Descriptor Visualization



KPI Details & Provenance Information

KPI Name

TotalGreenhouseGasEmissionsByWeightResultingFromPrintin

Context URI

http://sustainable.deri.ie/resource/report/context/context_2010

Unit

http://sustainable.deri.ie/measurementunits#kgco2e

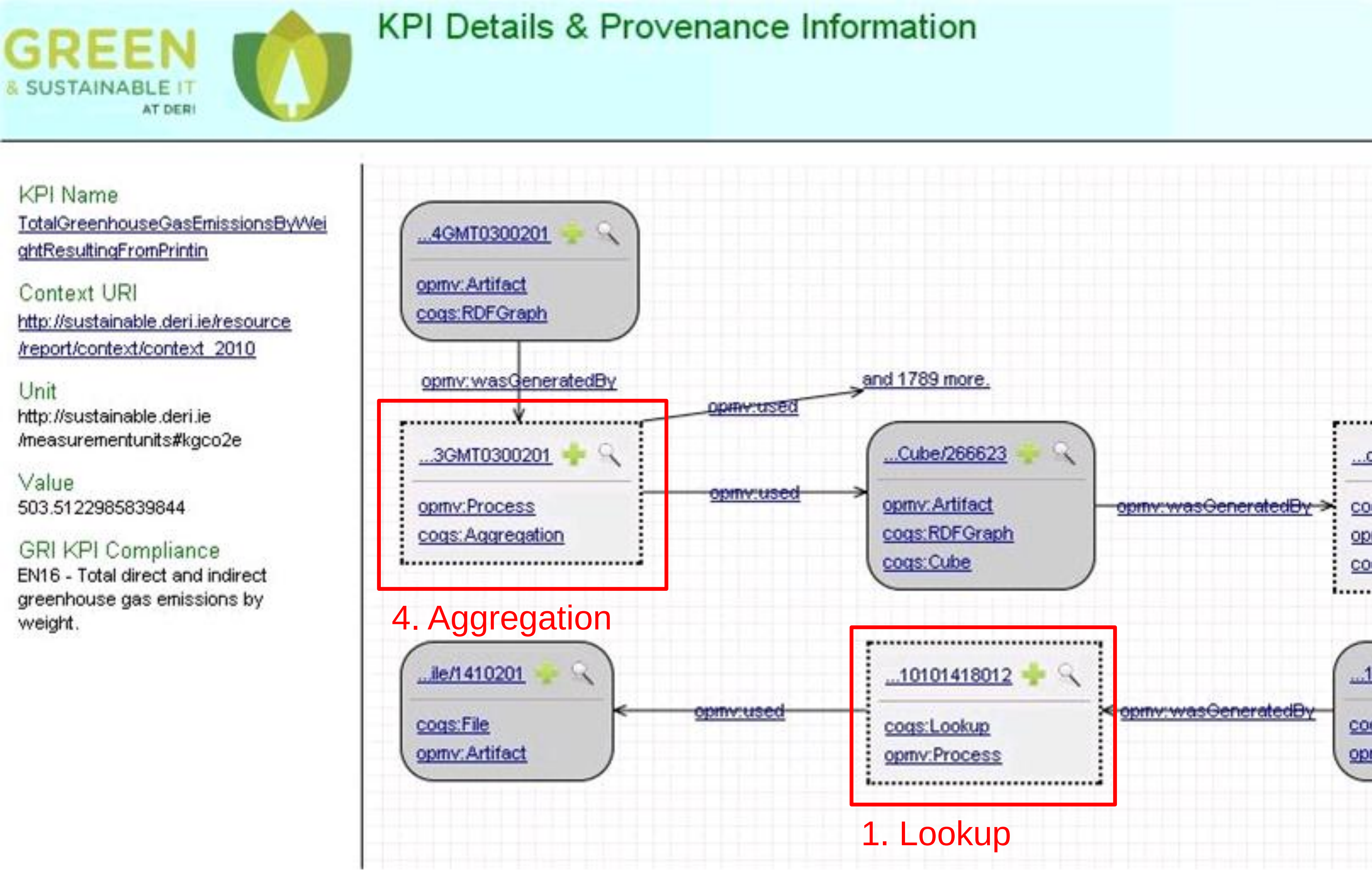
Value

503.5122985839844

GRI KPI Compliance

EN16 - Total direct and indirect greenhouse gas emissions by weight.

Case Study – Provenance Descriptor Visualization



KPI Details & Provenance Information

KPI Name
TotalGreenhouseGasEmissionsByWeightResultingFromPrintin

Context URI
http://sustainable.deri.ie/resource/report/context/context_2010

Unit
http://sustainable.deri.ie/measurementunits#kgco2e

Value
503.5122985839844

GRI KPI Compliance
EN16 - Total direct and indirect greenhouse gas emissions by weight.

4. Aggregation

1. Lookup

Case Study – Possible Queries

- OPMV
  - What are the data artifacts, processes and agents behind this data value?
  - When and how long were the processes executed?
- OPMV + Cogs
  - How *long* did all lookups take?
  - What *scripts* have been used to transform the data into RDF?
  - To which values *constant factors* have been applied?
  - Which *aggregation functions* were used to calculate this indicator?

Outline

- Motivation & Problem
- Gap of ETL Descriptions
- Interoperable ETL Provenance Model
- Case Study
- **Conclusions**

Conclusions

Cogs

Provenance representation from an ETL perspective

Semantic interoperability across different ETL applications

Usability and ontological commitment

- Evaluation in small case study
- For a full evaluation of interoperability benefits model needs to be adopted in provenance-aware ETL applications.
- Starting point: Provenance-aware Google Refine using Cogs.

Conclusions

Cogs

Provenance representation from an ETL perspective

Semantic interoperability across different ETL applications

Usability and ontological commitment

Thanks!